USE OF MEL FREQUENCY CEPSTRAL COEFFICIENTS FOR AUTOMATIC PATHOLOGY DETECTION ON SUSTAINED VOWEL PHONATIONS: MATHEMATICAL AND STATISTICAL JUSTIFICATION

R. Fraile, N. Sáenz-Lechón, J.I. Godino-Llorente, V. Osma-Ruiz, P. Gómez-Vilda

EUIT Telecomunicación. Circuits & Systems Engineering Department
Carretera de Valencia Km.7, 28031 Madrid (SPAIN)

ABSTRACT

This paper presents a justification for the use of MFCC parameters in automatic pathology detection on speech. While such an application has produced good results up to now, only partial explanations to this good performance had been given before. The herein exposed explanation consists of an interpretation of the mathematical transformations involved in MFCC calculation and a statistical analysis that confirms the conclusions drawn from the theoretical reasoning.

1. INTRODUCTION

Automatic detection of pathologies on speech has been traditionally realised through the analysis of distortion and noise measurements taken from records of sustained vowels [1]. However, recently alternative approaches have also been taken in two directions: using running text [1] and different parameterization schemes, such as those based on Mel-frequency Cepstral Coefficients (MFCC) [2]. While the use of this Mel-cepstral analysis produces low error rates, to authors' knowledge a complete justification for its use other than the empirical results has not been given yet, though the same authors have approached the problem in [3] and [4] and, previously, other authors have proven the presence of noise information in the speech cepstrum [5].

In this paper, a justification for the use of MFCC for pathology detection on speech records is provided. This consists in both a theoretical interpretation of the mathematical transformations involved in MFCC computation (section 2) and a statistical analysis of the parameters extracted from records belonging to a commercial database (section 3). In the statistical analysis, both parametric and non-parametric approaches have been taken. While the parametric analysis permits the use of analytic statistical tools, the non-parametric analysis allows confirming the results of the parametric analysis avoiding the model assumptions.

This research was carried out within projects funded by the Ministry of Science and Technology of Spain (TEC2006-12887-C02) and the Universidad Politécnica de Madrid (AL06-EX-PID-033). It has also been realised within the framework of European COST action 2103.

2. MATHEMATICAL FORMULATION OF CEPSTRAL COEFFICIENTS

The presence of pathologies in speech is closely related to signal variability, hence the traditional use of distortion measurements. The need for detecting such variability leads to the convenience of employing short-time techniques for speech processing. A mathematical framework for short-time processing of speech is provided in [6]. According to it, if a discrete-time speech signal $x[n]$ is segmented in speech frames $g_p[n]$ of length L, being p the frame index, then the short-time Discrete Fourier Transform (stDFT) of each frame may be written as:

$$S_p(k) = \sum_{n=0}^{L-1} g_p[n] \cdot e^{-j \cdot \frac{2\pi n}{N_{DFT}} \cdot k} \quad (1)$$

where N_{DFT} is the number of points of the DFT and k is the index of the DFT elements ($k = 0 \dots N_{DFT} - 1$).

2.1. Short-time Cepstrum

The short-time cepstrum can be derived from the stDFT as follows [6]:

$$c_p[q] = \frac{1}{N_{DFT}} \sum_{k=0}^{N_{DFT}-1} \log |S_p(k)| \cdot e^{j \cdot \frac{2\pi k}{N_{DFT}} \cdot q} \quad (2)$$

The use of cepstrum in the assessment of pathological voices is supported by two arguments: on the one hand, cepstrum analysis is appropriate for estimating the noise level of the voice signal [5] and, on the other hand, for the case of sustained vowels, the variability of the glottal waveform can also be easily detected from cepstral parameters [4]. A cepstral set of parameters calculated from a smoothed spectrum is proposed in [4]. This smoothing consists in the convolution in frequency domain of the stDFT with a set of triangular filters, much like in the case of MFCC [6]:

$$S'_p(i) = \sum_{f_k \in I_i} \left(1 - \frac{|f_k - i \cdot \Delta f/2|}{\Delta f/2}\right) \cdot |S_p(k)| \quad (3)$$

where $I_i = [\Delta f \cdot (i-1)/2, \Delta f \cdot (i+1)/2]$ and f_k is the frequency associated to $S_p(k)$. The number of filters M depends on the choice of Δf. Consequently: $i = 1 \ldots M$, and (2) becomes:

$$c'_p[q] = \frac{1}{M+1} \sum_{k=1}^{M} \log \left| S'_p(k) \right| \cdot \cos\left(\frac{\pi k}{M+1} \cdot q \right) \quad (4)$$

If the value of Δf is chosen to be higher than the fundamental frequency of the voice, then the pitch information is lost in the modified cepstrum (4). However, this fact does not significantly affect its pathology detection capability [4]. Nevertheless, up to now, no work on the optimisation of Δf (or M, equivalently) has been reported.

2.2. Short-time MFCC

Another approach for speech parametrisation in the cepstral domain is MFCC calculation. This approach is similar to the previous one in that it includes spectrum smoothing prior to the transformation into cepstral domain. However, it differs in that the spectrum smoothing is done in the Mel-frequency domain, hence resulting in a set of narrower filters for low frequencies and wider for high frequencies. Namely, the following frequency transformation is applied [7]:

$$f_k^m = 2595 \cdot \log_{10}\left(1 + \frac{f_k}{700}\right) \quad (5)$$

and the smoothing is done in the transformed domain:

$$\widetilde{S}_p(i) = \sum_{f_k^m \in I_i^m} \left(1 - \frac{\left| f_k^m - F^m \cdot \frac{i}{M+1} \right|}{\Delta f^m / 2} \right) \cdot |S_p(k)| \quad (6)$$

where $I_i^m = \left[F^m \cdot \frac{i-1}{M+1}, F^m \cdot \frac{i+1}{M+1} \right]$, M is the number of filters, $\Delta f^m = \frac{2}{M+1} \cdot F^m$ corresponds to Δf and F^m is the maximum frequency in Mel domain.

The number of Mel-band filters is commonly chosen to be: $M = \lfloor 3 \cdot \log f_s \rfloor$, but, again, the impact of the selection of M on the performance of the whole pathology detector has not been assessed up to now. After spectrum smoothing, the return to the linear frequency scale can be done considering that the central frequencies of the smoothing filters are $fc_i = 700 \cdot 10^{\frac{F^m \cdot i}{2595 \cdot (M+1)}}$. Then (4) becomes:

$$c''_p[q] = \frac{1}{M+1} \sum_{k=1}^{M} \log \left| \widetilde{S}_p(k) \right| \cdot \cos\left(\frac{2\pi f c_k}{f_s} \cdot q \right) \quad (7)$$

A graphical representation of both kinds of cepstral coefficients obtained after spectrum smoothing, namely *modified cepstrum* (4) and *Mel-band cepstrum* (7), is depicted in figure 1. It can be noticed that the use of Mel-band filters instead of the lineally spaced filters allows a reduction on the length of cepstrum, hence a dimensionality reduction, while keeping most of the information of the first cepstral coefficients. This is due to the Mel-band filters providing a more detailed description of the low frequency bands of speech and, on the opposite, averaging the less significant high frequency bands. Nevertheless, (7) is not the common way to express cepstral coefficients obtained from Mel-band spectrum smoothing. Instead, once the spectrum has been smoothed, the frequency transformation is not reversed, usually. This gives as a result what are commonly known as MFCC:

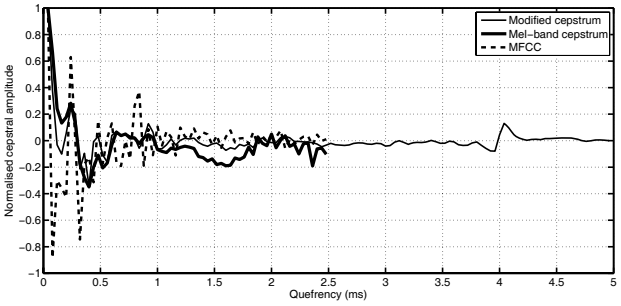

Figure 1. Modified cepstrum, Mel-band cepstrum and MFCC; all of them correspond to a speech record containing a sustained vowel $/ah/$.

$$c'''_p[q] = \frac{1}{M+1} \sum_{k=1}^{M} \log \left| \widetilde{S}_p(k) \right| \cdot \cos\left(\frac{\pi k}{M+1} \cdot q \right) \quad (8)$$

though a frequency-shifted version is more frequently used [7]:

$$c_p^{IV}[q] = \frac{1}{M+1} \sum_{k=1}^{M} \log \left| \widetilde{S}_p(k) \right| \cdot \cos\left(\frac{\pi (k-0.5)}{M+1} \cdot q \right) \quad (9)$$

In fact, both (8) and (9) give similar results. A representation of the MFCC obtained using (8) is also plot in figure 1. It can be observed that the Mel-frequency transformation produces in cepstral domain somewhat of a quefrency compression of the slower component of cepstrum, thus reducing the quefrency interval containing significant values of cepstrum.

As a conclussion to this section, it can be said that the capability of cepstrum to carry information of both noise level of sustained vowels [5] and their associated glottal waveform [4] is kept when the Mel-band filters are used for spectrum smoothing. Moreover, the Mel-frequency transformation produces in cepstral domain a compression of the quefrency axis that, in principle, should be useful in providing some reduction in the size of speech feature vectors.

3. PERFORMANCE ANALYSIS

Within this section, the performance in pathology detection of each one of the three above-described approaches for speech parametrisation in cepstral domain is analysed. A three-step analysis is presented. The first step consists in assessing the relevance of each individual parameter for pathology detection using the Fisher linear discriminant

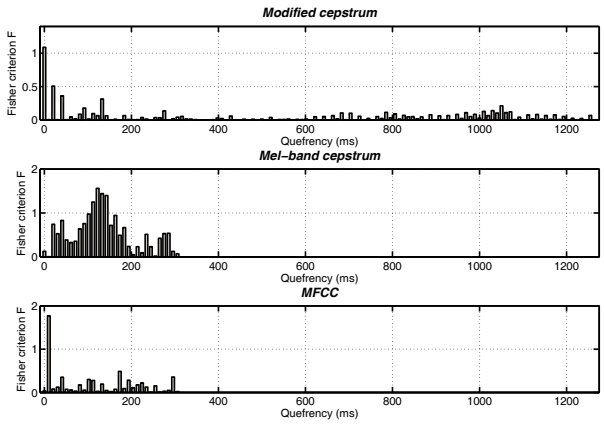

Figure 2. Value of the Fisher criterion for each cepstral parameter: modified cepstrum (4) -up-, Mel-band cepstrum (7)-middle- and MFCC (8) -down-.

criterion. The second step is based on the typical "Gaussian" assumption for the distribution of the speech segments in the feature vector space. The last part consists in studying the performance of non-parametric classifiers so as to confirm the results obtained in the second part.

For all three parts, the well-known Kay speech record database is used [8]. More specifically, a subset containing 53 normal and 173 pathological records, each consisting of a sustained phonation (1-3 s long) of the vowel $/ah/$ [9]. This subset covers a wide variety of voice disorders and the distribution of speakers is balanced in age and gender. The sampling rate of speech records has been made equal to 25 kHz, while coding has a resolution of 16 bits. For short-time processing, each speech record has been split in 20 ms frames and cepstral analysis has been performed for each individual frame.

3.1. Relevance of individual parameters

The first step in the analysis has consisted in assessing the relevance of each individual cepstral parameter for pathology detection. More specifically, the parameters given by (4) for $\Delta f = 200\ Hz$ and (7) and (8) for $M = 31$ have been analysed by means of the Fisher linear discriminant [10]. The greater values of the discriminant correspond to the greater relevances of the parameters for detection.

Figure 2 shows the calculated values of the Fisher discriminant for all the cepstral coefficients of each scheme. The obtained results confirm the observations made in the previous section. Firstly, the spectrum smoothing using Mel-band filters allows both a reduction in the length of cepstrum and an increase in the relevance of computed parameters for pathology detection. This can be concluded from comparison of the top and middle graphs in figure 2. Secondly, observation of both the middle and bottom graphs indicates that keeping the Mel-frequency transformation when passing from the spectral to the cepstral domain further increases the relevance of the first cepstral parameters while reducing that of the rest.

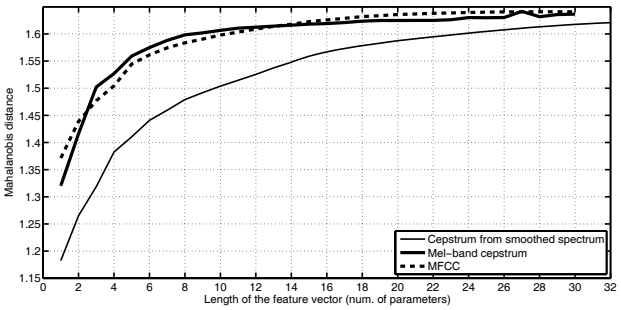

Figure 3. Mahalanobis distance as a function of the length of the feature vector for the three types of cepstral analysis under study.

3.2. Parametric analysis of feature vectors

In general, the analysis of individual parameters, such as the one reported before, is not fully significant when such individual parameters belong to multidimensional feature vectors. This is due to the joint relevance of parameters not depending only on their individual relevances, but also on the correlation between their values. It is hard to find an statistical test that considers both individual relevance and correlation for arbitrarily distributed feature vectors. For this reason, such distribution has been assumed Gaussian and, based on it, the Mahalanobis distance [10] has been taken as a performance indicator for each combination of parameters. Its value is directly related to the performance of the feature vector, that is, the greater the Mahalanobis distance associated to a feature vector, the less detection error rate should be expected from that vector.

In the herein reported analysis, the values of the Mahalanobis distance have been calculated over the abovementioned database. The selection of parameters in order to compose the feature vectors has been realised as follows: first, the parameter yielding the highest Fisher discriminant value has been chosen and, afterwards, an iterative process has been run in order to, at each step, add to the feature vector the parameter that gave the greatest increase in the Mahalanobis distance. Results are plot in figure 3.

The relation of the calculated Mahalanobis distance with the length of the feature vector indicates that similar performances are to be expected from the three sets of cepstral parameters, but the use of Mel-band spectrum smoothing allows achieving that performance with shorter feature vectors. To give a figure, while the asymptotic behaviour of the MFCC and Mel-band cepstral vectors is reached with vector lengths between 15 and 20, lengths well above 26 are necessary for modified cepstral vectors. Looking at another aspect of the graph in figure 3, the performance of Mel-band cepstrum and MFCC is very similar, though MFCC provide better performance for very short feature vectors. This is coherent with previous results, since the first element of the MFCC vector is more significant than the first element of the Mel-band cepstrum.

The same approach including the evaluation of the Ma-

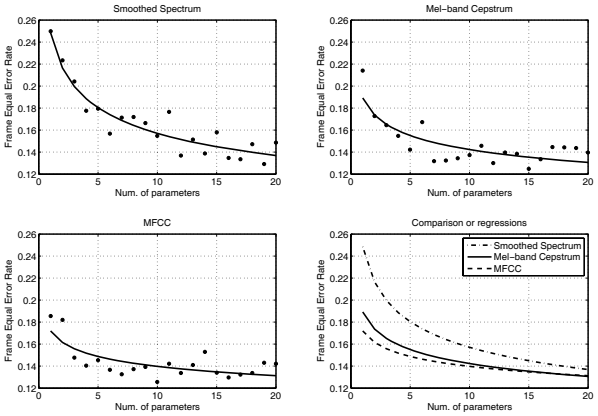

Figure 4. Comparison between the three schemes in terms of performance as the length of the feature vector increases. Potential regressions have been plotted so as to ease interpretation.

halanobis distance has been undertaken so as to assess the impact of the spectral filter width selection on the performance of each parameter set. For Mel-band cepstrum the option $M = 3 \cdot \log f_s = 31$ has been found to provide somewhat of an upper limit on the performance. For MFCC slight improvements might be achieved by increasing this value; however, the impact of increasing M has shown to be much more significant for $M < 31$ than for $M > 31$; thus $M = 31$ appears to be a sensible option too. For the case of modified cepstrum, the same trend occurs: for narrow filters (larger M) performance is better than for wider filters (lower M). Namely, $\Delta f = 100\ Hz$ and $\Delta f = 200\ Hz$ exhibit similar performace, while for $\Delta f > 200\ Hz$ the performance degrades.

3.3. Non-parametric analysis

In order to confirm the previous results, a non-parametric analysis has been realised. This consists in measuring the equal error rate (EER) achieved with each one of the three schemes when the classification of the feature vectors is carried out by a Multilayer Perceptron (MLP), which is a general purpose non-parametric classifier [10]. Dependence of the achieved EER on the length of the feature vector is depicted in figure 4. The selection of features for each case has been done according to the criterion based on the Mahalanobis distance and explained before. The graph is fully coherent with that of figure 3: MFCC and Mel-band cepstrum have very similar performances, with a slight advantage of MFCC for short vectors, whilst modified cepstrum needs longer feature vectors to achieve comparable performance.

4. CONCLUSION

Within this paper, it has been shown that MFCC are relevant parameters for automatic pathology detection. Mathematically, it has been shown that their computation allows to concentrate the noise information present in speech cepstrum in a few cepstral coefficients, namely those corresponding to the lowest quefrencies. Such conclusion has been confirmed by means of a statistical analysis. This analysis has also demonstrated that MFCC provide certain advantage over other similar cepstral analyses in terms of dimensionality reduction.

5. REFERENCES

[1] K. Umapathy, S. Krishnan, V. Parsa, and D. G. Jamieson, "Discrimination of pathological voices using a time-frequency approach," *IEEE Trans. on Biomed. Eng.*, vol. 52, no. 3, pp. 421–430, Mar 2005.

[2] J. I. Godino-Llorente and P. Gómez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Trans. on Biomed. Eng.*, vol. 51, no. 2, pp. 380–384, Feb 2004.

[3] J. I. Godino-Llorente, P. Gómez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters," *IEEE Trans. on Biomed. Eng.*, vol. 53, no. 10, pp. 1943–1953, Oct 2006.

[4] R. Fraile, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and P. Gómez-Vilda, "Use of cepstrum-based parameters for automatic pathology detection on speech. Analysis of performance and theoretical justification," in *Proc. of Biosignals 2008*, vol. 1, Jan 2008, pp. 85–91.

[5] P. J. Murphy and O. O. Akande, "Quantification of glottal and voiced speech harmonics-to-noise ratios using cepstral-based estimation," in *Proc. of the 3th Internat. Conf. on Non-Linear Speech Proces. (NOLISP'05)*, 2005, pp. 224–232.

[6] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals*. New York (USA): Macmillan Publishing Company, 1993.

[7] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition*. Englewood Cliffs (USA): Prentice-Hall, 1993.

[8] Kay Elemetrics Corp., "Disordered voice database." 1994.

[9] V. Parsa and D. G. Jamieson, "Identification of pathological voices using glottal noise measures," *Journal of Speech, Lang. and Hearing Research*, vol. 43, pp. 469–485, Apr 2000.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. New York (USA): John Wiley & Sons, 2001.